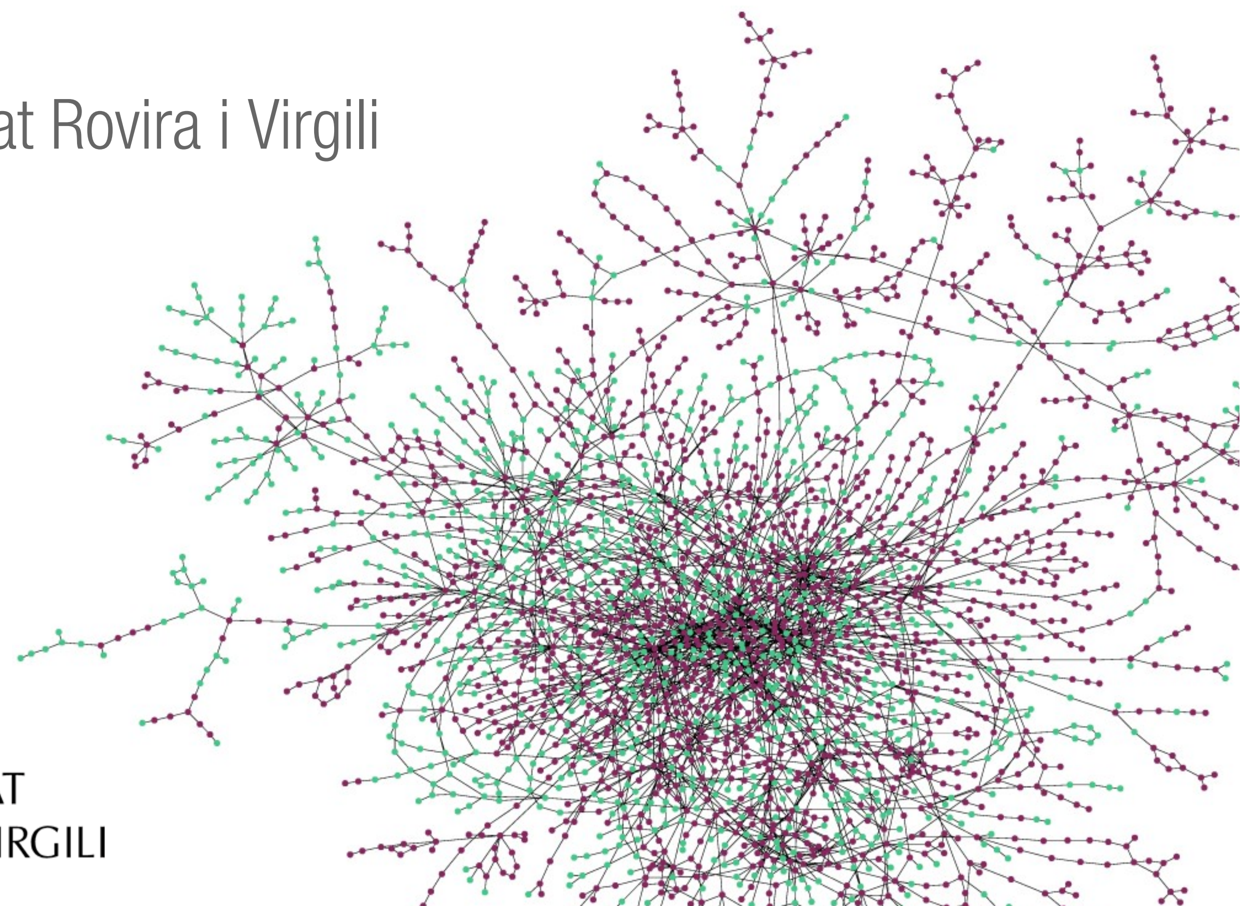
Statistical inference for the discovery of hidden interactions in complex networks

Roger Guimerà
ICREA and
Chemical Engineering, Universitat Rovira i Virgili

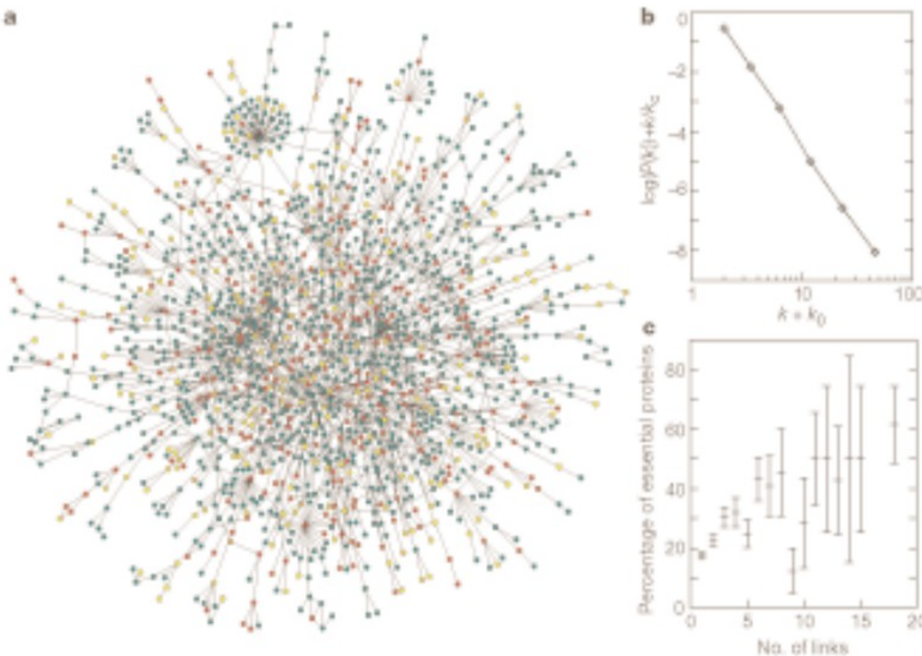
NetSci'13
Copenhagen, June 4, 2013



*iCrea
INSTITUCIÓ CATALANA DE RECERCA I ESTUDIS AVANÇATS

UNIVERSITAT ROVIRA I VIRGILI

One billion dollars to map the human proteome



Jeong, et al., *Nature* (2001)

naturenews

nature news home | news archive | specials | opinion | features | news blog | nature jo

comments on this story

ories by subject

Biotechnology
Business
Cell and molecular biology
Chemistry
Genetics
Health and medicine
Lab life
Technology

ories by keywords

Proteomics
Human Proteome Project
Human Proteome Organisation
Proteins
Human Genome Project

his article elsewhere

Published online 23 April 2008 | **452**, 920-921 (2008) | doi:10.1038/452920a

News

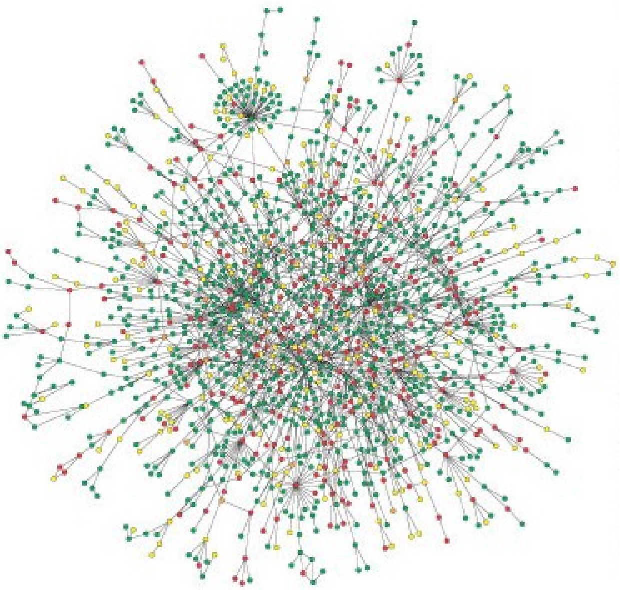
Biologists initiate plan to map human proteome

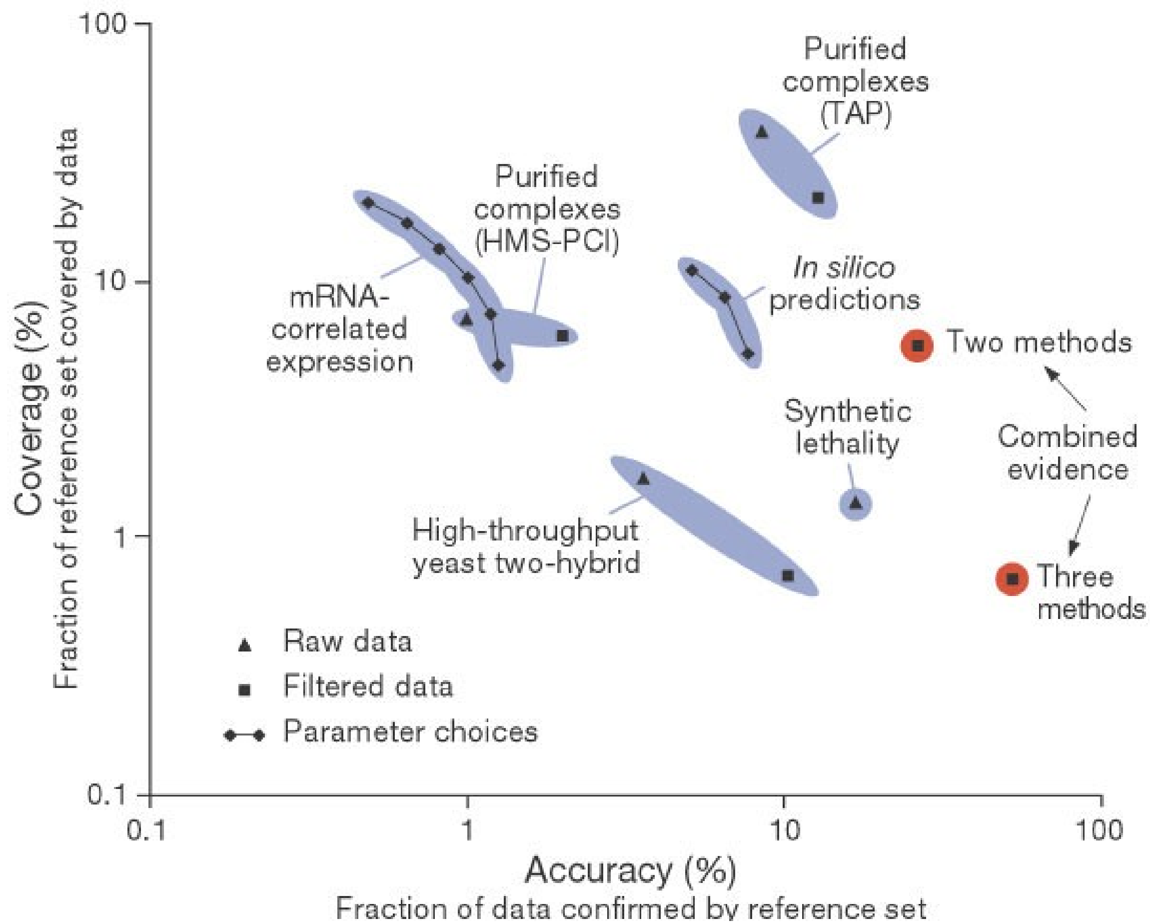
Project aims to characterize all human proteins.

Helen Pearson

Ambitious plans to catalogue and characterize all proteins in the human body — a Human Proteome Project — are being drawn up by a small group of researchers. But with a price tag of around US$1 billion, some question whether the organizers can raise enough money or momentum for such an undertaking.

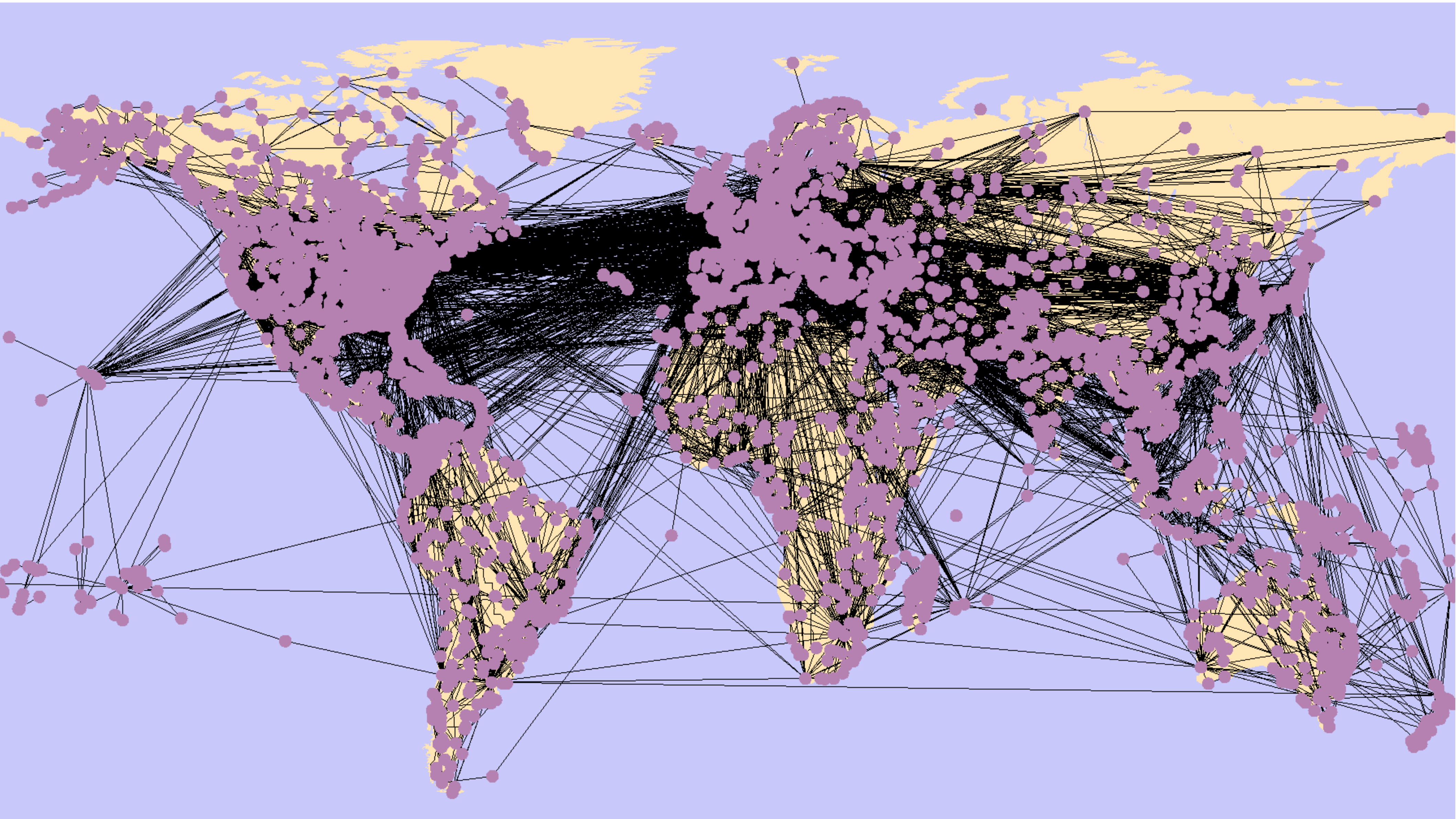
Accuracy and coverage are a concern for protein interaction (and most other) datasets



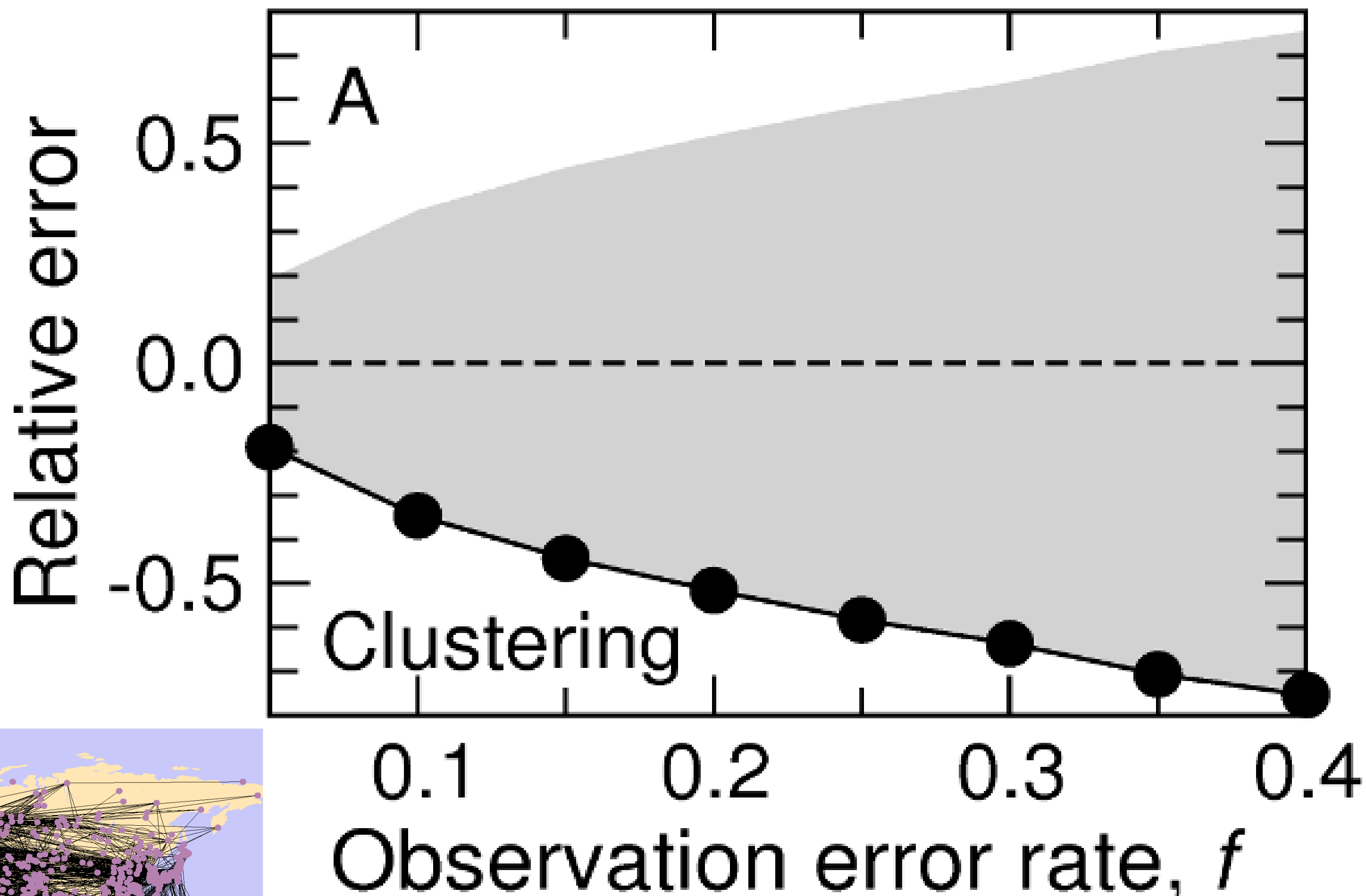


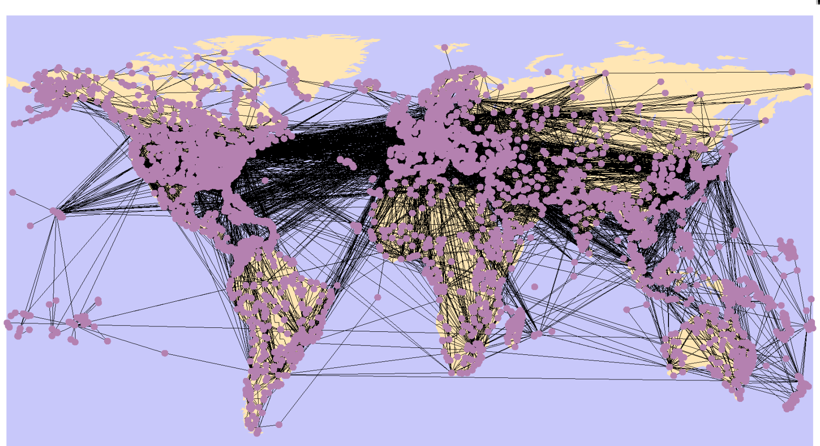
von Mering et al., *Nature* (2002)

All network data is subject to noise

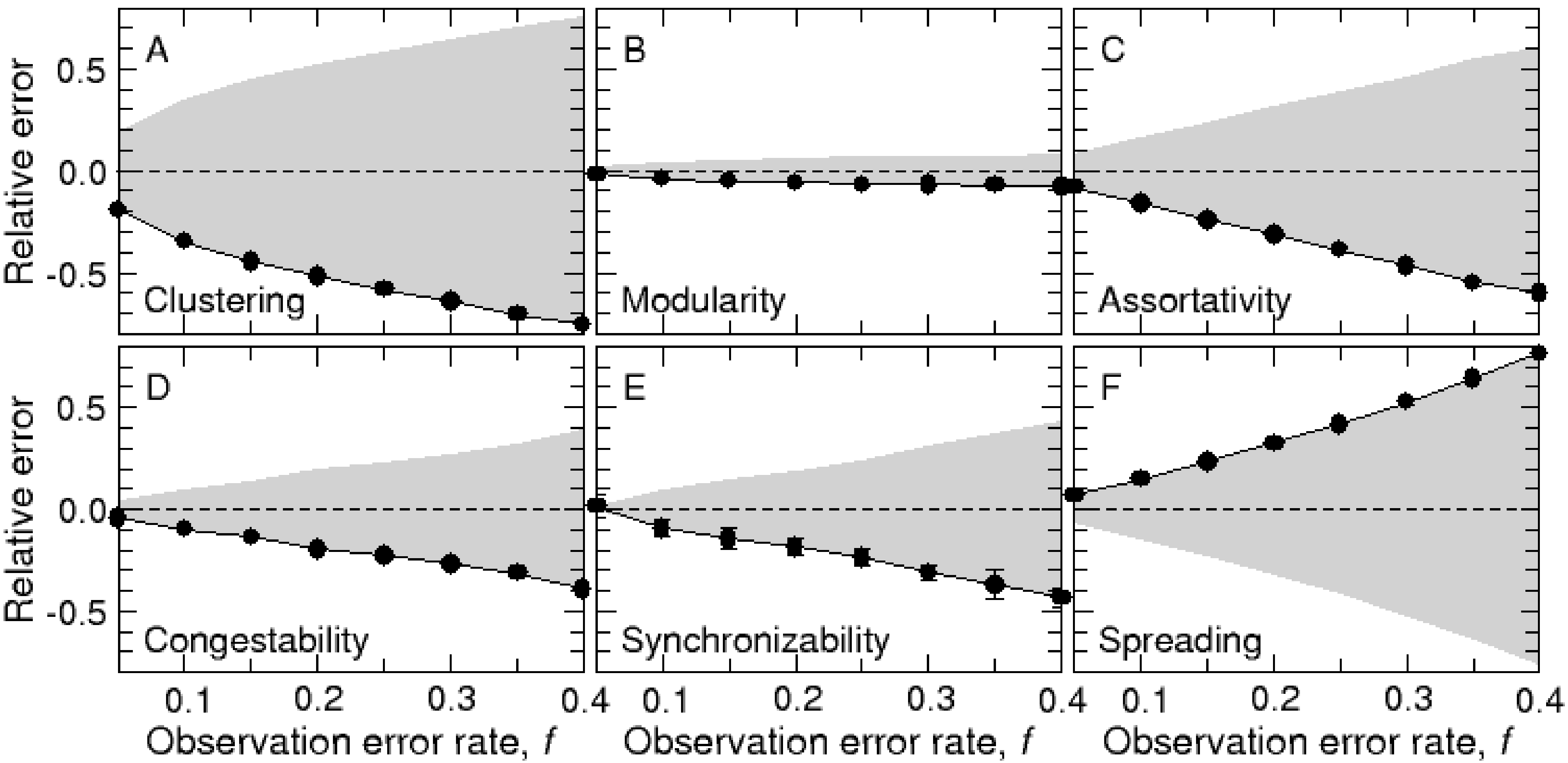


Network properties are often sensitive to even low error rates

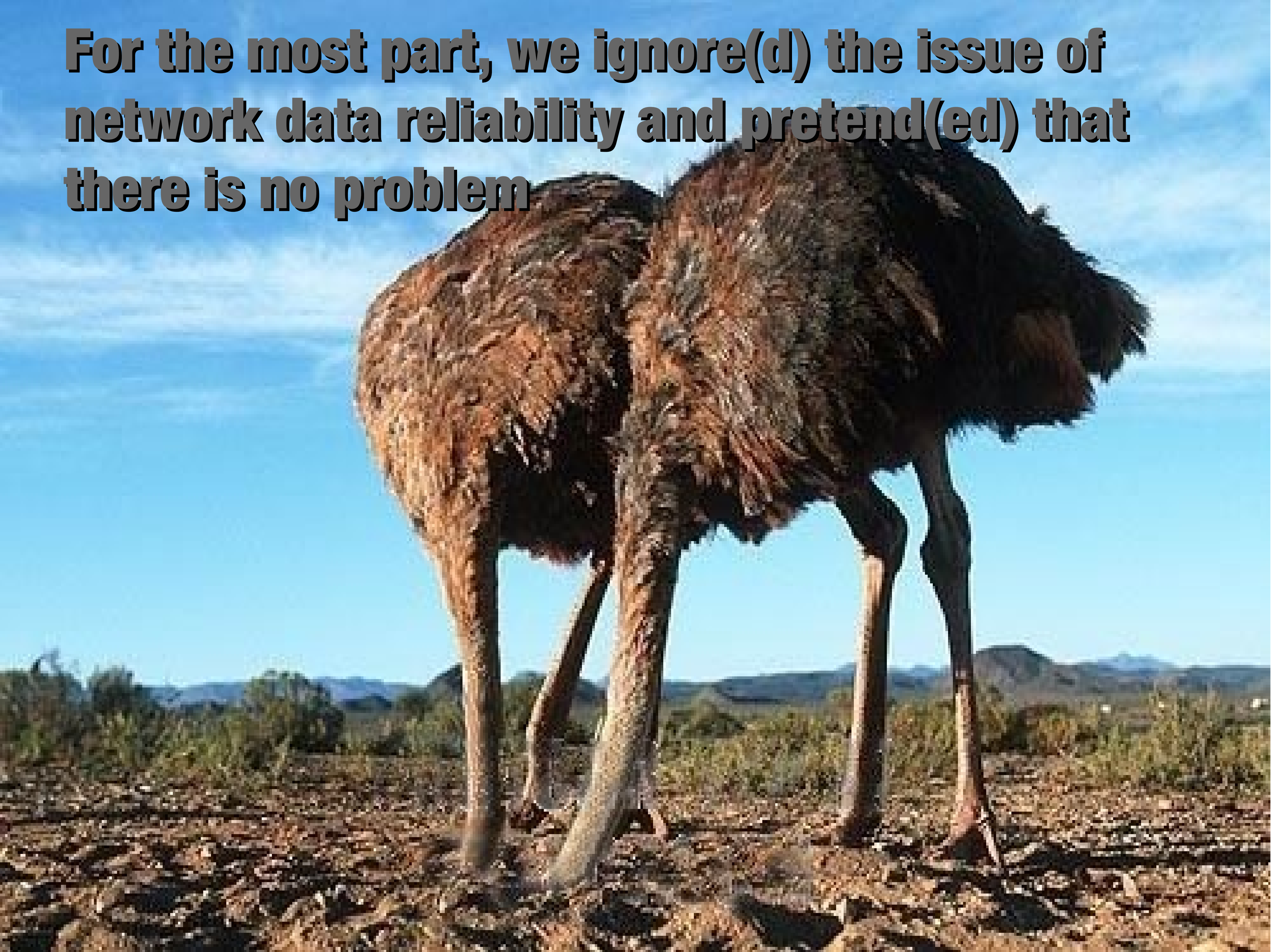




Network properties are often sensitive to even low error rates

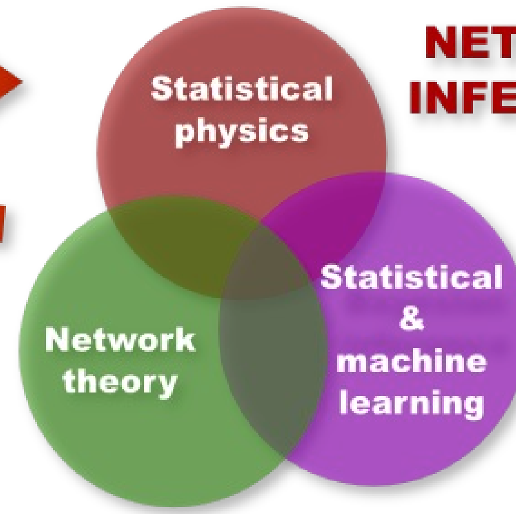


Guimera, Sales-Pardo, *PNAS* (2009)


For the most part, we ignore(d) the issue of network data reliability and pretend(ed) that there is no problem

Outline

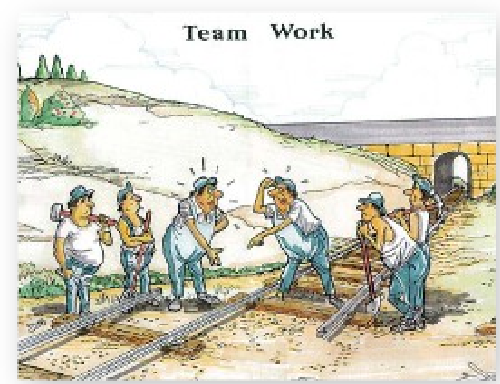
Can we help to clean up noisy network data?



NETWORK INFERENCE

Can we uncover unknown drug interactions?

Can we predict human decisions?



Can we predict conflict in small teams?

What is to be done?

- Given a single noisy observation of a network, determine:
  - *Missing interactions* Interactions that exist but are not captured in our observation of the system
  - *Spurious interactions* Interactions that do *not* exist but, for some reason, are included in our observation
- *Reconstruct the network*, so that our reconstruction has properties that are closer to the properties of the true network

What is to be done?

- Given a single noisy observation of a network, determine:
  - *Missing interactions* Interactions that exist but are not captured in our observation of the system
  - *Spurious interactions* Interactions that do *not* exist but, for some reason, are included in our observation
- *Reconstruct the network*, so that our reconstruction has properties that are closer to the properties of the true network
- But:
  - We want to be able to do this for arbitrary real networks about which we don't know anything
  - There seems to be a paradox in trying to identify what is wrong in a network observation–from *the network observation itself* !

There are two possible scenarios when in comes to solving the paradox

- Scenario 1: We *don't* have a clue about what the network should look like, or where does it come from (mechanistically or statistically):
  - We cannot do anything
- Scenario 2: We *do* have some ideas about the structure of the network:
  - We can formalize these ideas into a set of models
  - We can use the models to assess what is likely to be missing/wrong

The "reliability formalism"

- We assume our network is the outcome of an undetermined model M from a (potentially infinite) collection of models $\mathcal{M}$
- We observe a network A^O
- Given my observation A^O, what is the probability that a property X takes the value $X=x$ if we generate a new network (with the same model)?

$$p(X=x|A^O) = \int_{\mathcal{M}} dM\, p(X=x|M)\, p(M|A^O)$$
$$= \frac{\int_{\mathcal{M}} dM\, p(X=x|M)\, p(A^O|M)\, p(M)}{\int_{\mathcal{M}} dM\, p(A^O|M)\, p(M)}$$

- We call $p(X=x|A^O)$ the reliability of the $X=x$ measurement

Guimera, Sales-Pardo, *PNAS* (2009)

In particular, one can use the formalism to infer missing and spurious interactions

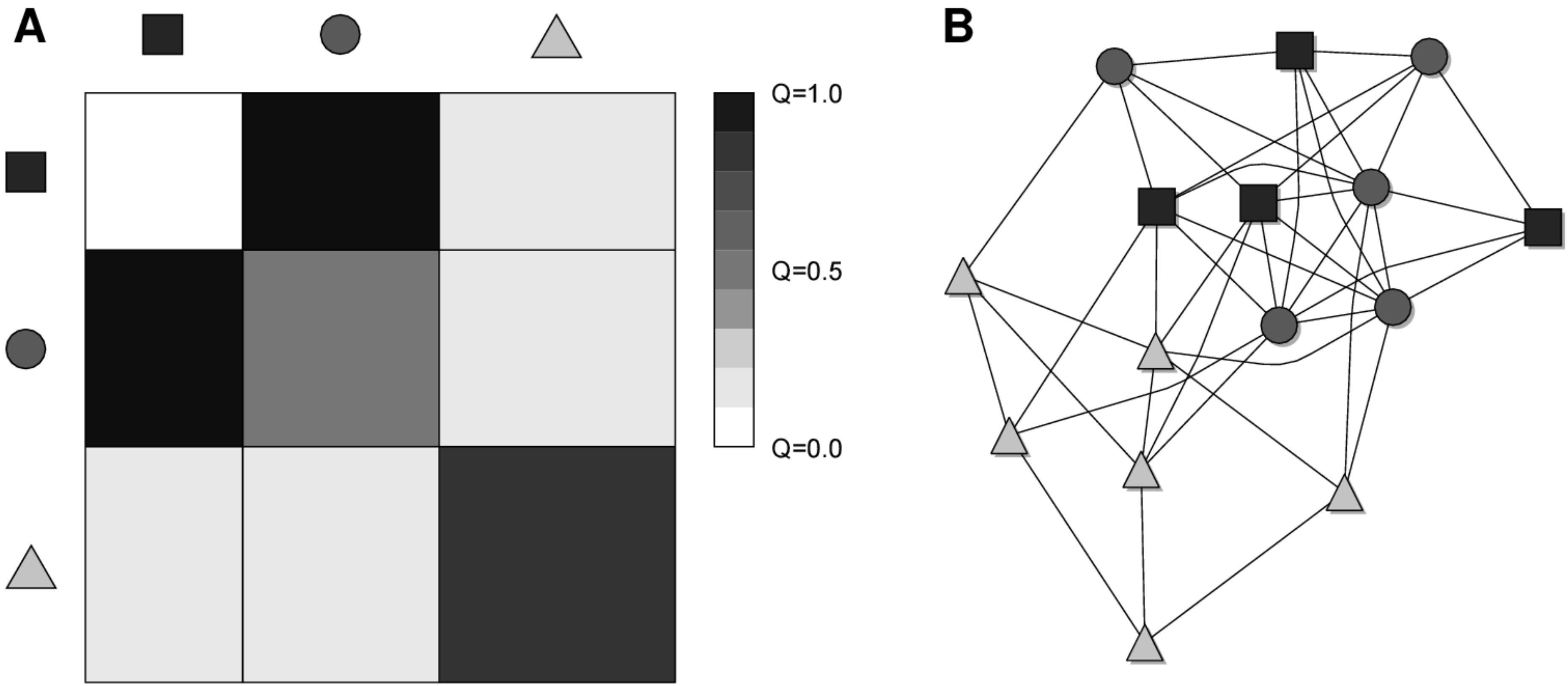
$$p(A_{ij}=1|A^O) = \frac{\int_{\mathcal{M}} dM\, p(A_{ij}=1|M)\, p(A^O|M)\, p(M)}{\int_{\mathcal{M}} dM\, p(A^O|M)\, p(M)}$$

- What property of networks is general enough that applies to *all* complex networks?
  - Broad (scale-free) connectivity distribution? No
  - Small world property? Yes—but no realistic/tractable model
  - Modularity? Group structure? YES

Clauset, Moore, Newman, *Nature* (2008)

Guimera, Sales-Pardo, *PNAS* (2009)

Stochastic block models (SBM) are *general*, *empirically grounded* and analytically *tractable*



- ➔ A stochastic block model is fully determined by a partition of the nodes into groups and the probabilities Q that a node in a group is connected to a node in any other group

White, Boorman, Breiger, *AJS* (1976)

Holland, Laskey, Leinhardt, *Soc. Networks* (1983)

Nowicki, Snijders, *JASA* (2001)

Stochastic block models (SBM) are *general*, *empirically grounded* and analytically *tractable*

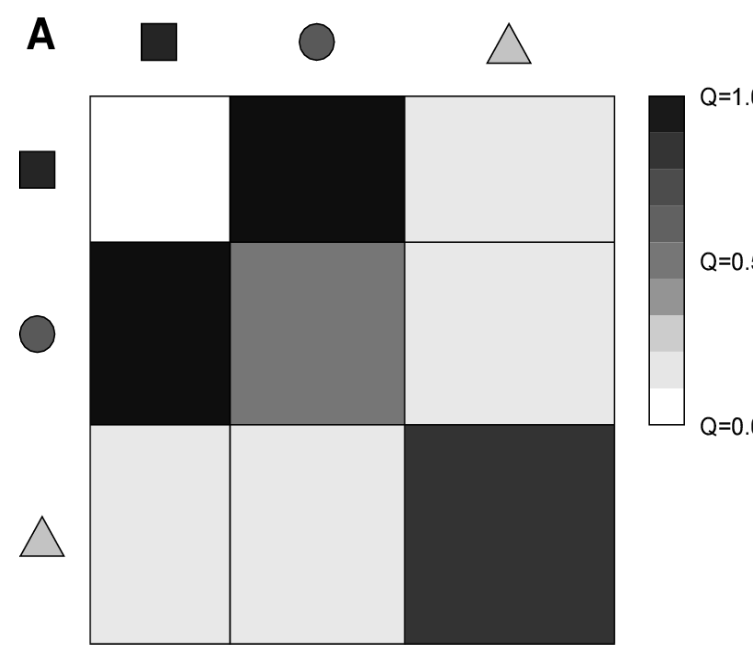
$$p(A_{ij}=1|A^O)=\frac{\int_{\mathcal{M}} dM\, p(A_{ij}=1|M)\, p(A^O|M)\, p(M)}{\int_{\mathcal{M}} dM\, p(A^O|M)\, p(M)}$$

$$p(A_{ij}=1|M)=Q_{\sigma_i\sigma_j}$$

$$p(A^O|M)=\prod_{\alpha\leq\beta} Q_{\alpha\beta}^{n^1_{\alpha\beta}}(1-Q_{\alpha\beta})^{n^0_{\alpha\beta}}$$

$$p(M)=\text{constant}$$

$$\int_{\mathcal{M}} dM \to \sum_{P\in\mathcal{P}} \prod_{\alpha\leq\beta}\left(\int_0^1 dQ_{\alpha\beta}\right)$$



Guimera, Sales-Pardo, *PNAS* (2009)

The link reliability is an ensemble average over all possible partitions of the nodes into groups

- In the end, the reliability of a link is

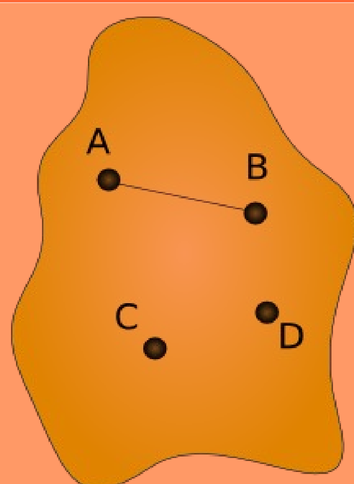
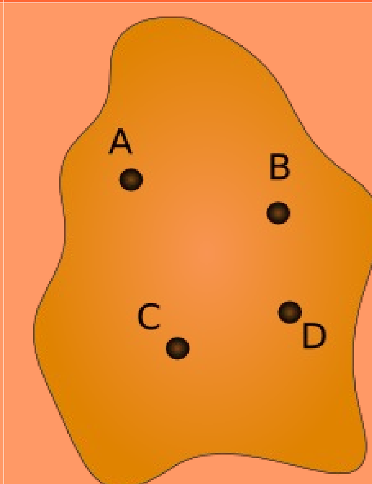
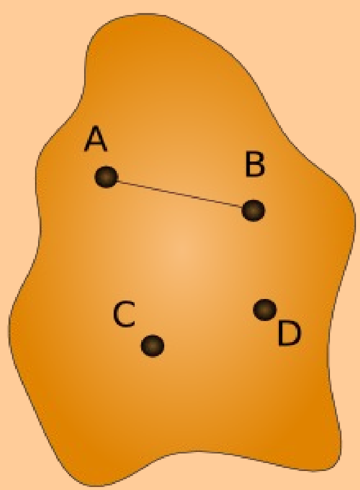
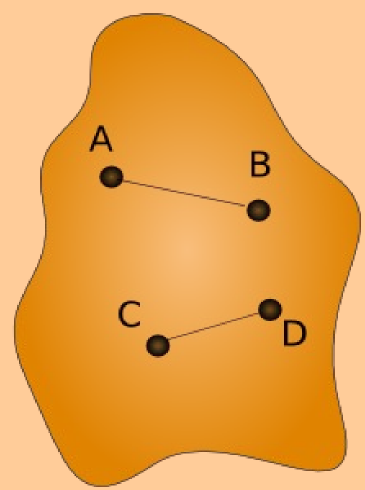
$$p(A_{ij} = 1 | A^O) = \frac{1}{Z} \sum_{P \in \mathcal{P}} \left(\frac{n^1_{\sigma_i \sigma_j} + 1}{n^0_{\sigma_i \sigma_j} + n^1_{\sigma_i \sigma_j} + 2} \right) \exp[-\mathcal{H}(\mathcal{P})]$$

- Where:

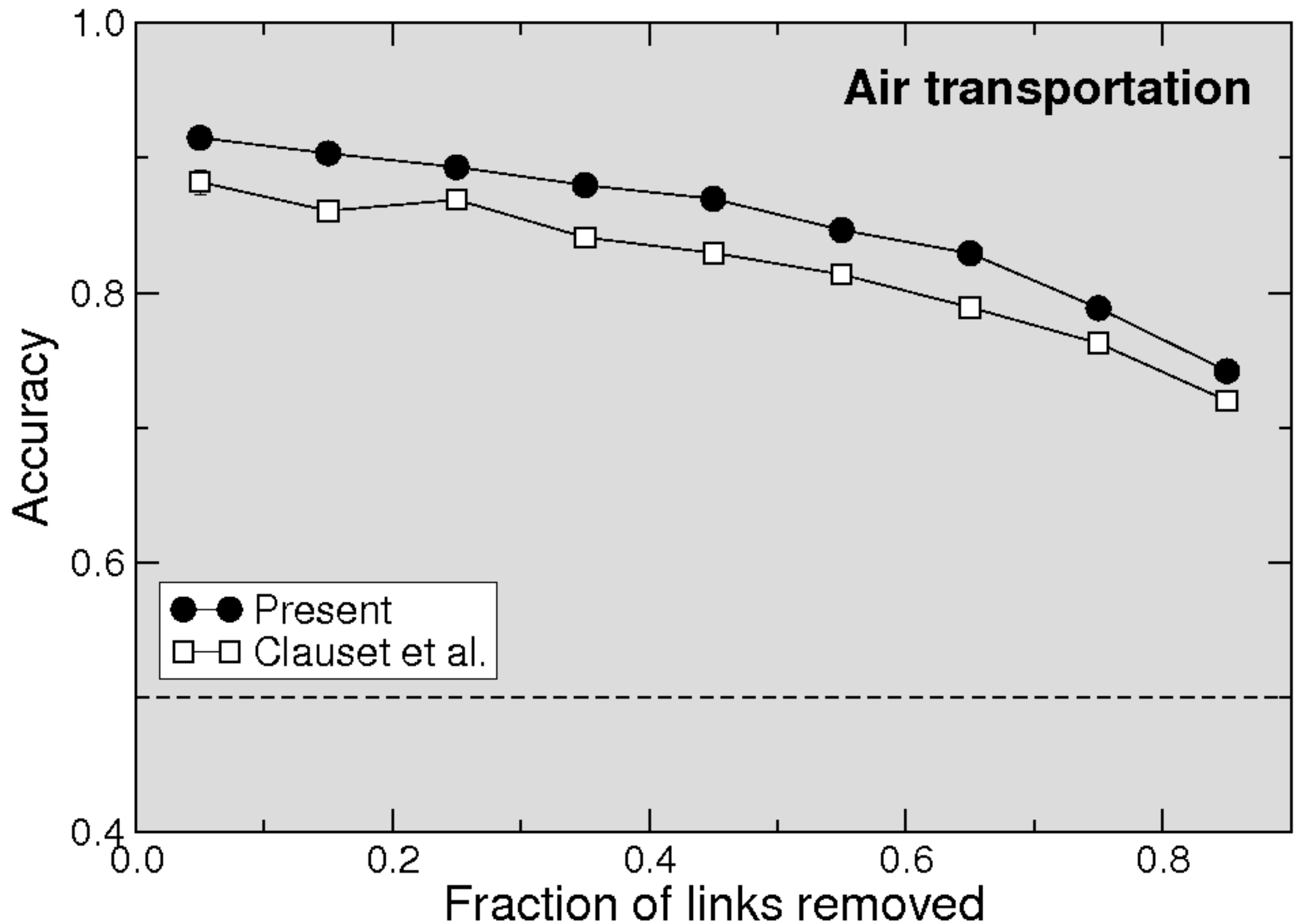
$$\mathcal{H}(\mathcal{P}) = \sum_{\alpha \leq \beta} \left[\ln(n_{\alpha\beta} + 1)! - \ln(n^0_{\alpha\beta})! - \ln(n^1_{\alpha\beta})! \right]$$

Guimera, Sales-Pardo, *PNAS* (2009)

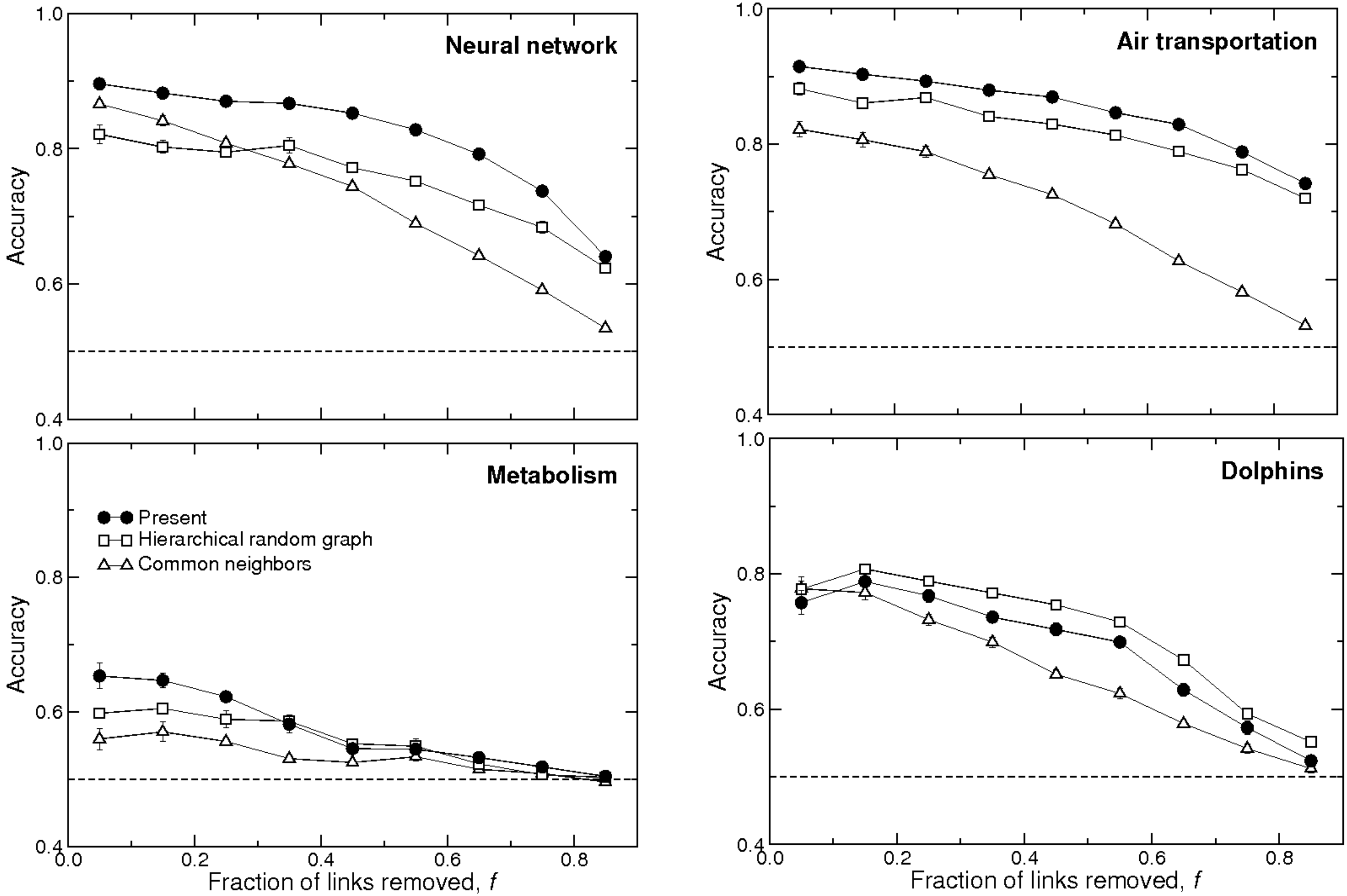
We test our algorithm to see if it can identify missing and spurious interactions in real networks

	True network	Observed network	Test
Missing interactions			How often is AB more reliable than CD?
Spurious interactions			How often is CD less reliable than AB?

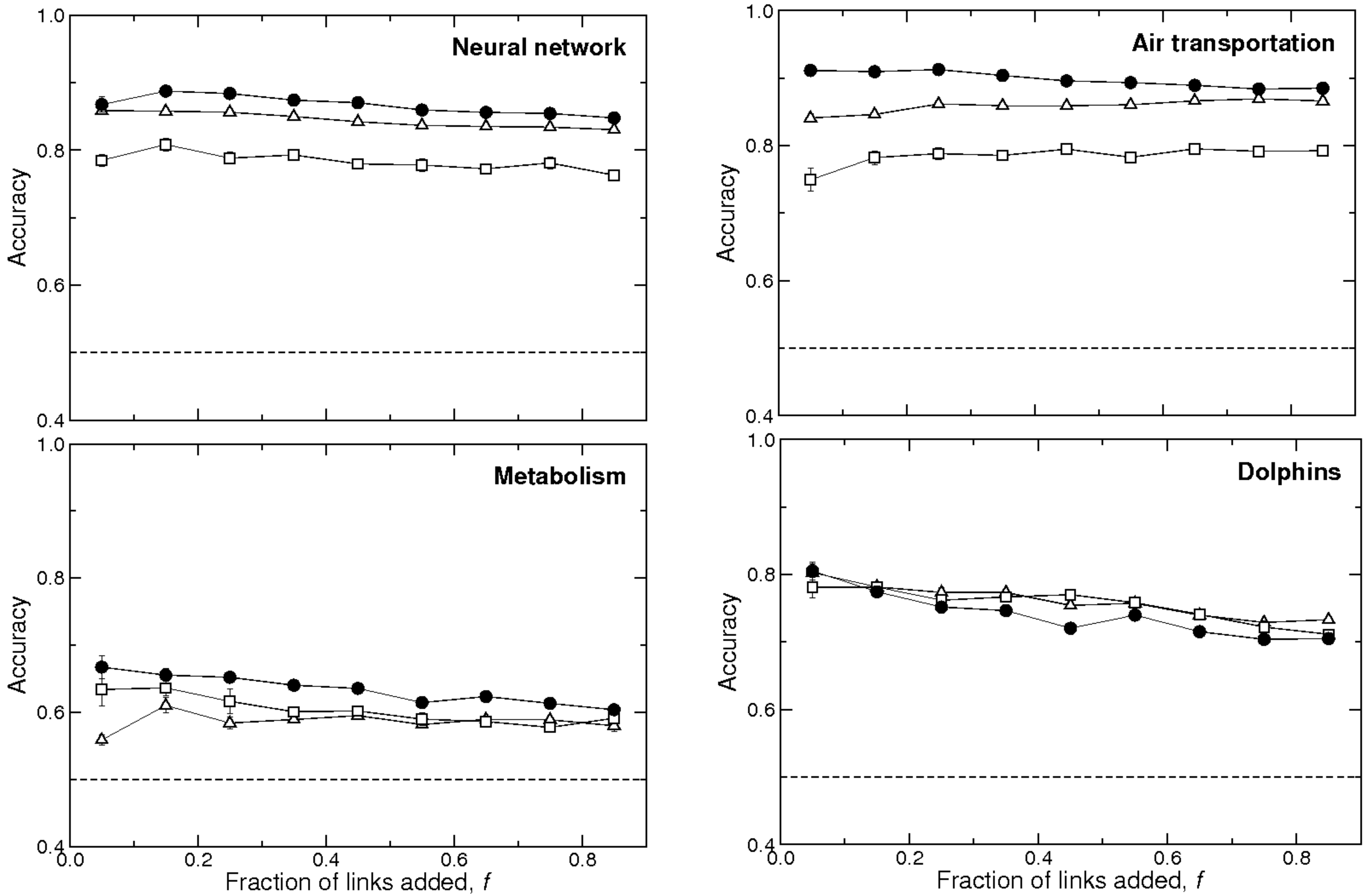
Our approach accurately recovers missing interactions



Our approach accurately recovers missing interactions



Our approach accurately recovers spurious interactions



Wonkish interlude I: H, module identification, maximum likelihood block models and all that

$$p(A_{ij} = 1|A^O) = \frac{1}{Z} \sum_{P \in \mathcal{P}} \left(\frac{n^1_{\sigma_i \sigma_j} + 1}{n^0_{\sigma_i \sigma_j} + n^1_{\sigma_i \sigma_j} + 2} \right) \exp[-\mathcal{H}(\mathcal{P})]$$

- What is this "energy"?

$$\mathcal{H}(P) = -\ln p(P|A^O)$$

- Therefore, the partition that minimizes **this** energy is the most likely given the data (except for priors, degree correction of the block model...):
  - More appropriate "modularity" function
  - No need to play with the number of groups
  - No over-fitting

Wonkish interlude II

$$\text{Unipartatite unweighted:} \quad \mathcal{H}(\mathcal{P}) = \sum_{\alpha \leq \beta} \left[\ln(n_{\alpha\beta}+1)! - \ln(n^0_{\alpha\beta})! - \ln(n^1_{\alpha\beta})!\right]$$

$$\text{Unipartite weighted:} \quad \mathcal{H}(\mathcal{P}) = \sum_{\alpha \leq \beta} \left[\ln(n_{\alpha\beta}+K-1)! - \sum_{k=1}^{K} \ln(n^k_{\alpha\beta})!\right]$$

$$\text{Bipartite weighted:} \quad \mathcal{H}(\mathcal{P}_{\mathcal{U}}, \mathcal{P}_{\mathcal{I}}) = \sum_{\alpha, \beta} \left[\ln(n_{\alpha\beta}+K-1)! - \sum_{k=1}^{K} \ln(n^k_{\alpha\beta})!\right]$$

Guimera, Sales-Pardo, *PNAS* (2009)

Guimera, Sales-Pardo, *PLOS ONE* (2011)

Guimera, Llorente, Moro, Sales-Pardo, *PLOS ONE* (2012)

Rovira-Asenjo, Gumi, Sales-Pardo, Guimera, *in press* (2013)

Reconstructing a network is more complicated than just adding missing interactions and removing spurious interactions

- Challenges:
  - We don't know how many links need to be added and removed
  - Links cannot be added and removed independently of each other

We define a network reliability

➔ The reliability of a network is

$$p(A|A^O) = \frac{1}{Z} \sum_{P \in \mathcal{P}} f(A; A^O, P) \exp[-\mathcal{H}(\mathcal{P})]$$

Guimera, Sales-Pardo, *PNAS* (2009)

The *network reconstruction* is the most reliable network

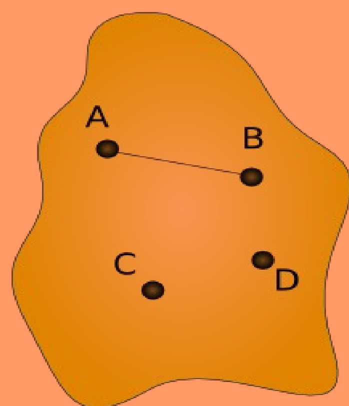
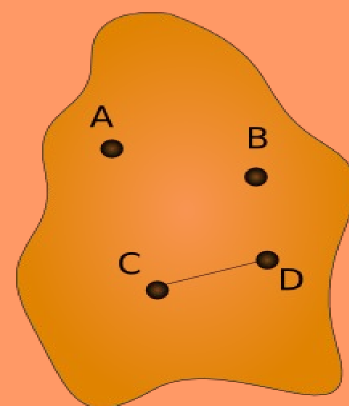
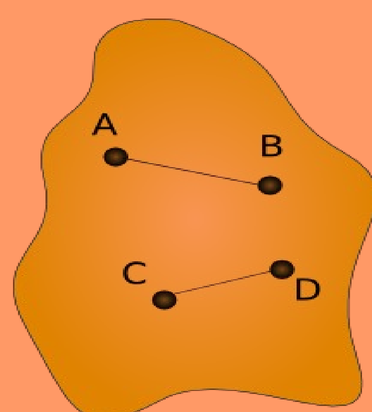
- The reliability of a network is

$$p(A|A^O) = \frac{1}{Z} \sum_{P \in \mathcal{P}} f(A; A^O, P) \exp[-\mathcal{H}(\mathcal{P})]$$

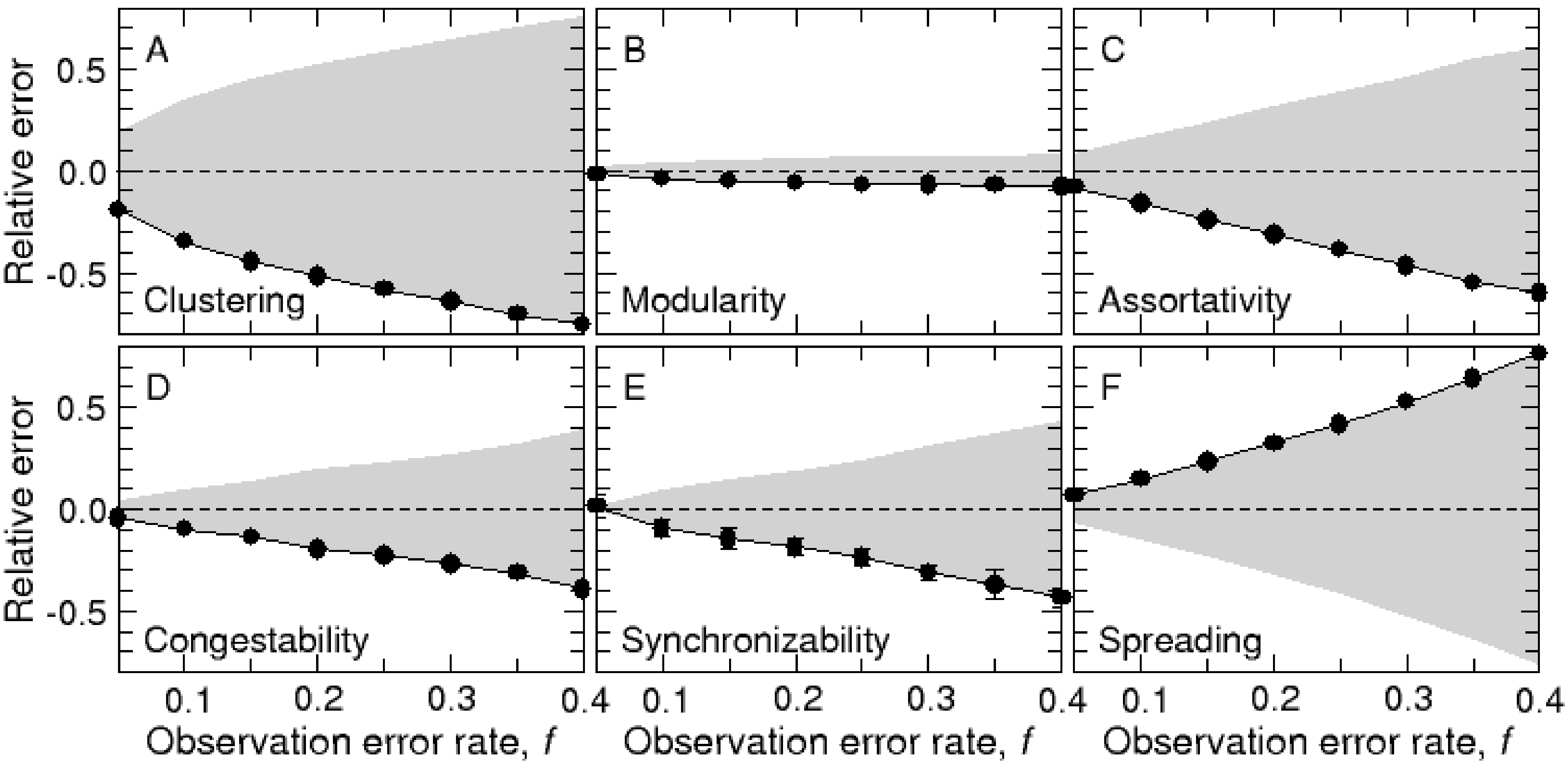
- The reconstruction A^R is the network that maximizes this probability
- We obtain A^R using uphill search

Guimera, Sales-Pardo, *PNAS* (2009)

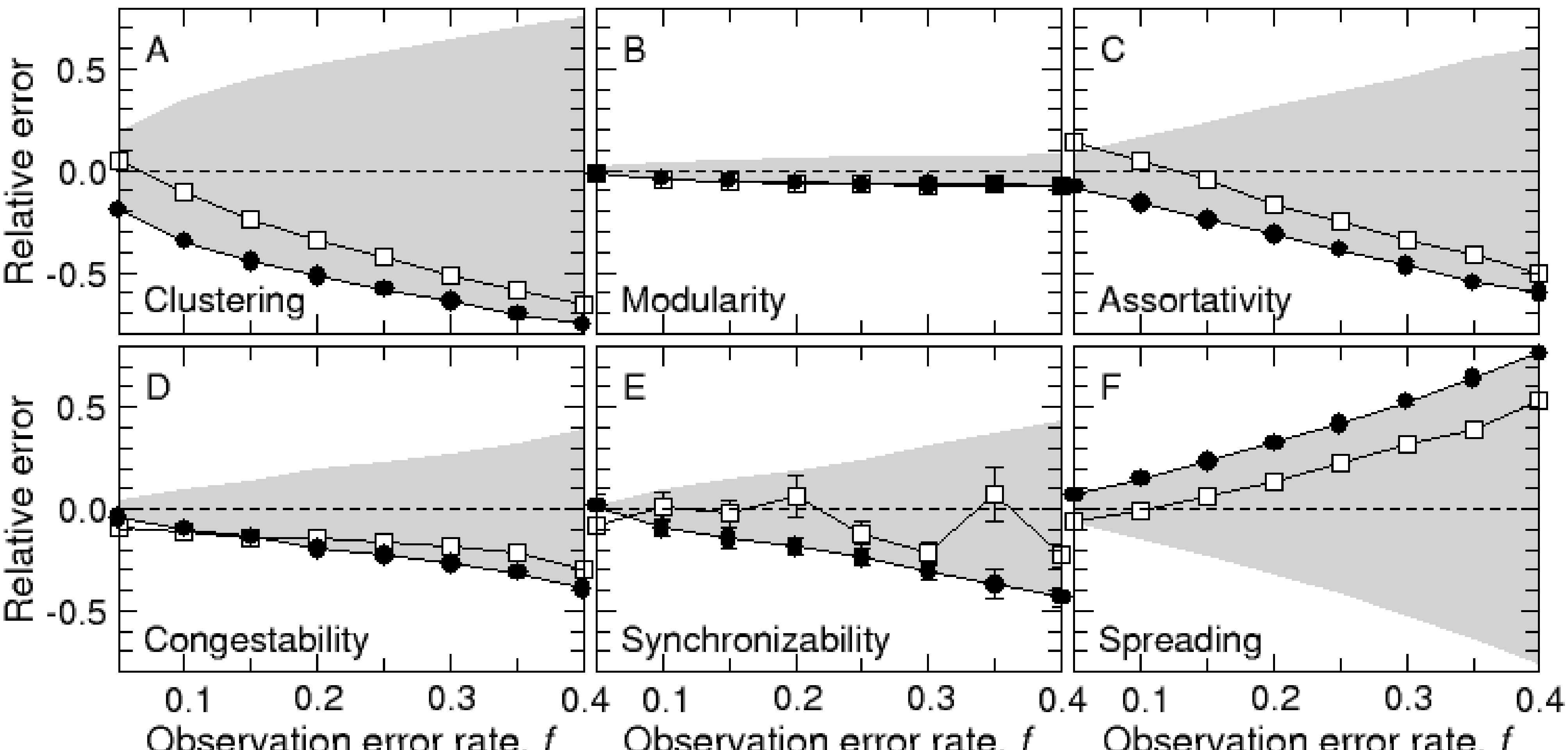
We can test what is the effect of random errors in our network observations

<table>
<tr><th></th><th>True network</th><th>Observed network</th><th>Test</th></tr>
<tr><td>Random errors</td><td>
</td><td>
</td><td>How do network properties change?</td></tr>
<tr><td></td><td></td><td>Reconstructed network</td><td></td></tr>
<tr><td></td><td></td><td>
</td><td>How do network properties change?</td></tr>
</table>

Network reconstructions provide better estimates of global network properties than the observations themselves



Network reconstructions provide better estimates of global network properties than the observations themselves

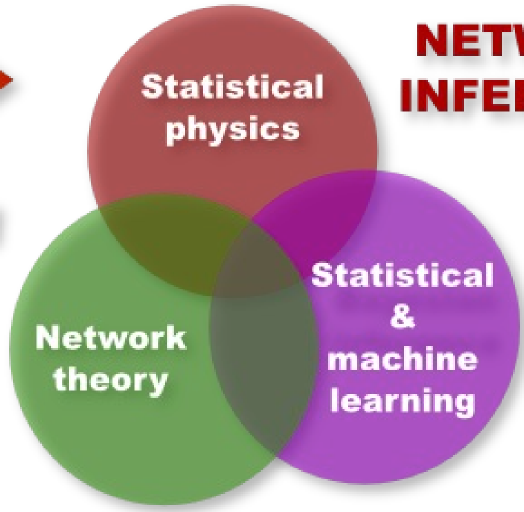


Guimera, Sales-Pardo, *PNAS* (2009)

Outline

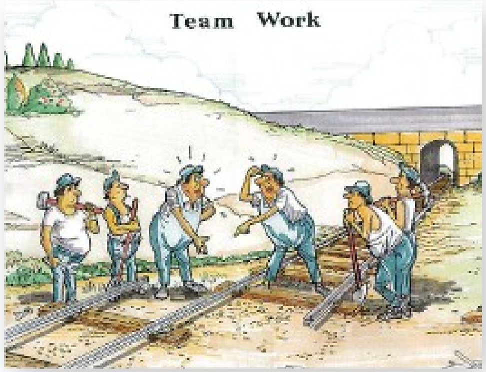
Can we help to clean up noisy network data?

NETWORK INFERENCE



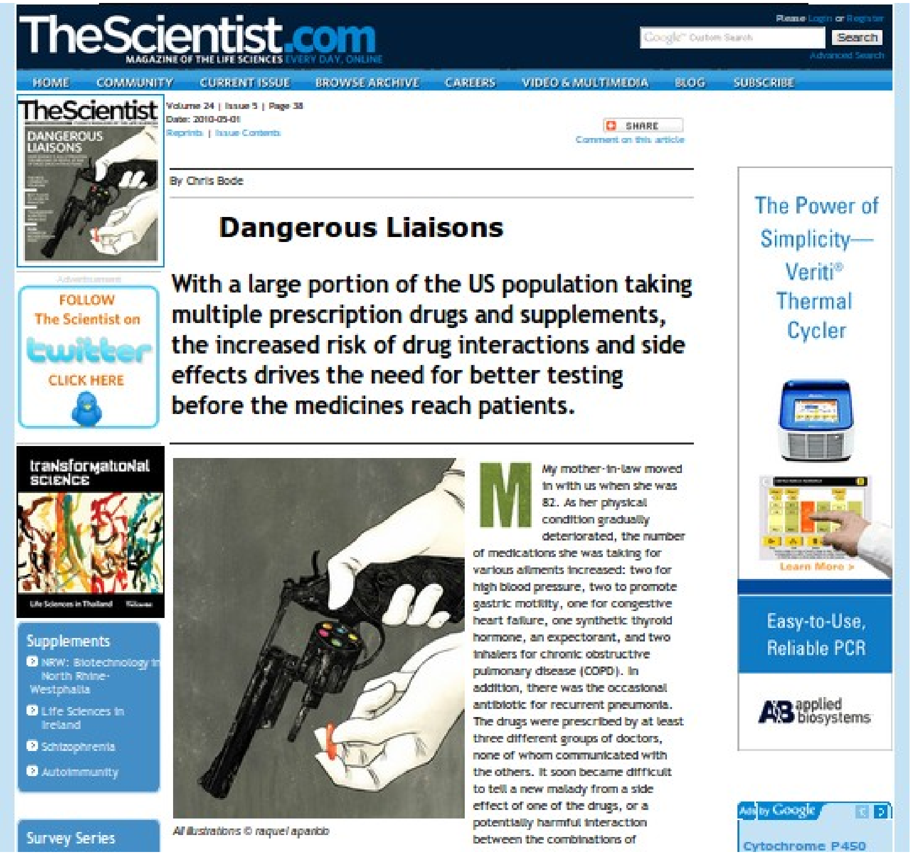
Can we uncover unknown drug interactions?

Can we predict human decisions?



Can we predict conflict in small teams?

The challenge of discovering novel drug-drug interactions



TheScientist.com
MAGAZINE OF THE LIFE SCIENCES EVERY DAY, ONLINE

HOME COMMUNITY CURRENT ISSUE BROWSE ARCHIVE CAREERS VIDEO & MULTIMEDIA BLOG SUBSCRIBE

Volume 24 | Issue 5 | Page 38
Date: 2010-05-01
Reprints | Issue Contents

By Chris Bode

Dangerous Liaisons

With a large portion of the US population taking multiple prescription drugs and supplements, the increased risk of drug interactions and side effects drives the need for better testing before the medicines reach patients.

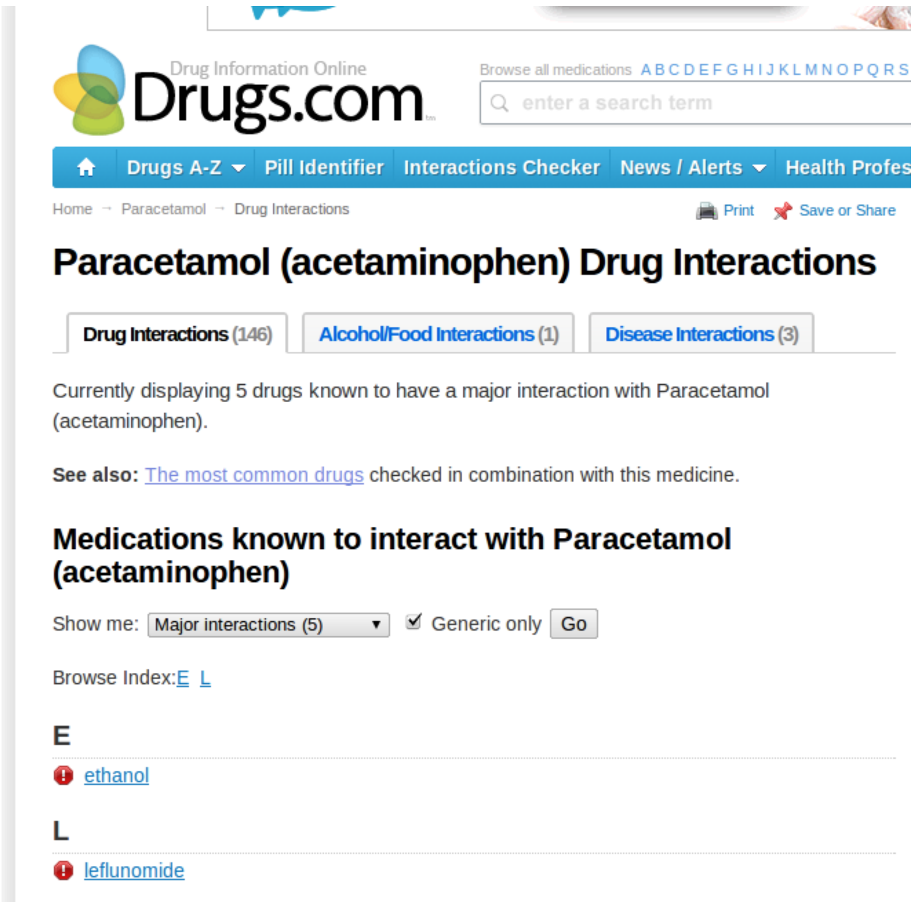
My mother-in-law moved in with us when she was 82. As her physical condition gradually deteriorated, the number of medications she was taking for various ailments increased: two for high blood pressure, two to promote gastric motility, one for congestive heart failure, one synthetic thyroid hormone, an expectorant, and two inhalers for chronic obstructive pulmonary disease (COPD). In addition, there was the occasional antibiotic for recurrent pneumonia. The drugs were prescribed by at least three different groups of doctors, none of whom communicated with the others. It soon became difficult to tell a new malady from a side effect of one of the drugs, or a potentially harmful interaction between the combinations of

All illustrations © raquel aparicio

- Twenty-nine percent [of U.S. population aged 57-85] used at least 5 prescription medications concurrently.
- Overall, 4% of individuals were potentially at risk of having a major drug-drug interaction.

Qato et al. *JAMA* (2008)

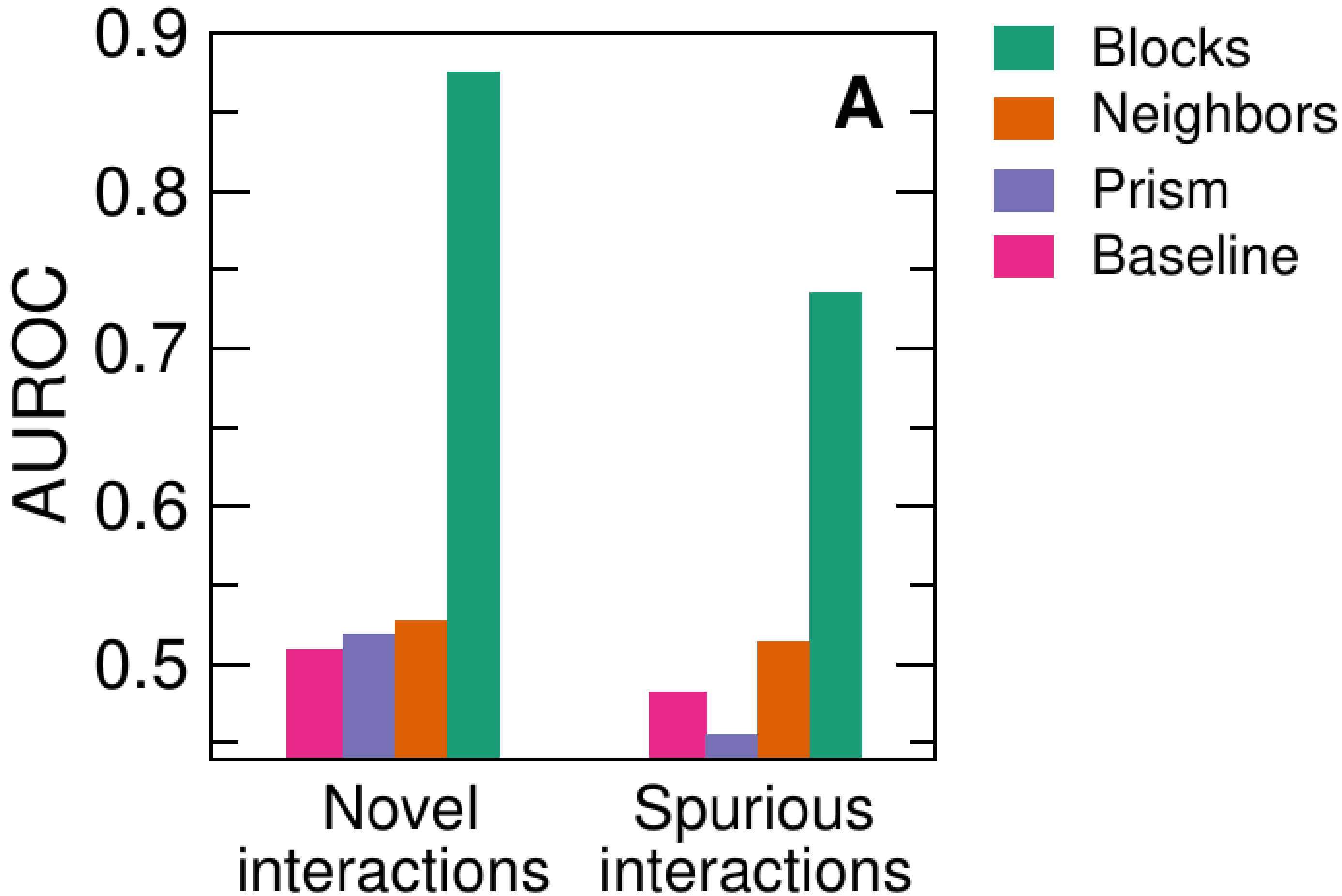
Can we predict which severe drug interactions will be dded to / removed from a database?



- Two snapshots of the drug-interaction database available at *drugs.com*:
  - May 10th, 2010
  - February 22nd, 2012
- Between the snapshots:
  - 1349 interactions added
  - 165 interactions removed

Guimera, Sales-Pardo, *submitted* (2013)

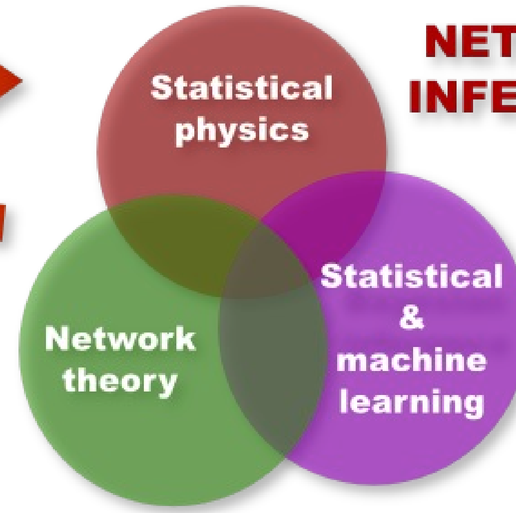
We can predict which severe drug interactions will be removed from and added to a database



Guimera, Sales-Pardo, *submitted* (2013)

Outline

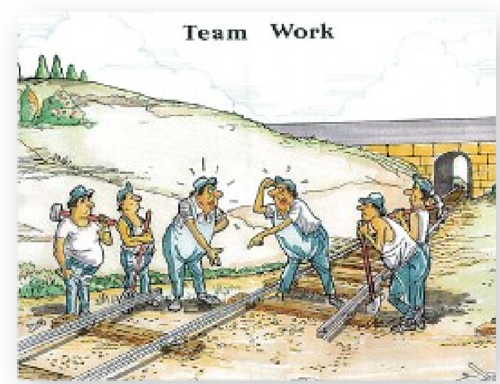
Can we help to clean up noisy network data?



NETWORK INFERENCE

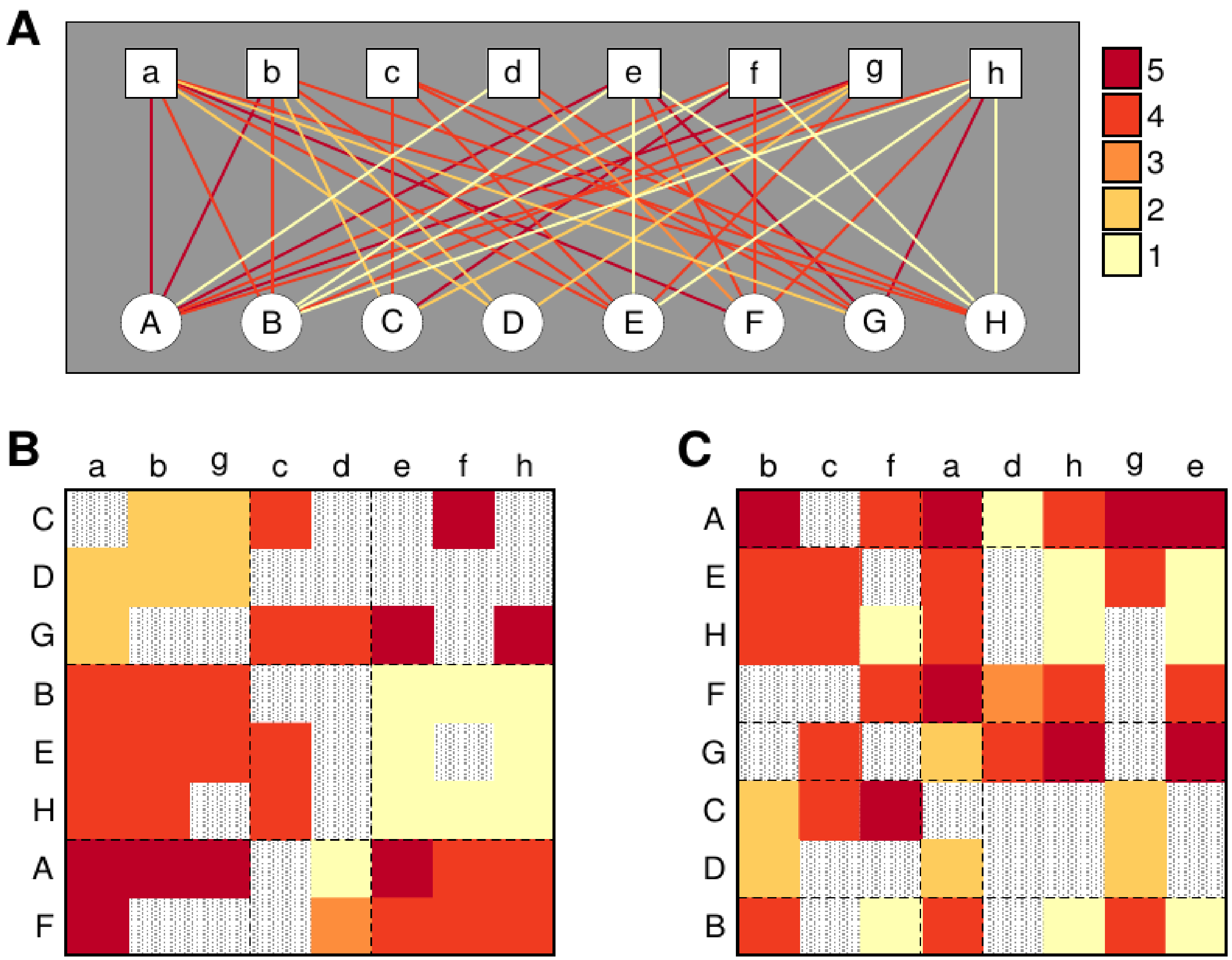
Can we uncover unknown drug interactions?

Can we predict human decisions?



Can we predict conflict in small teams?

Predicting human preferences can be reformulated as a problem of network inference



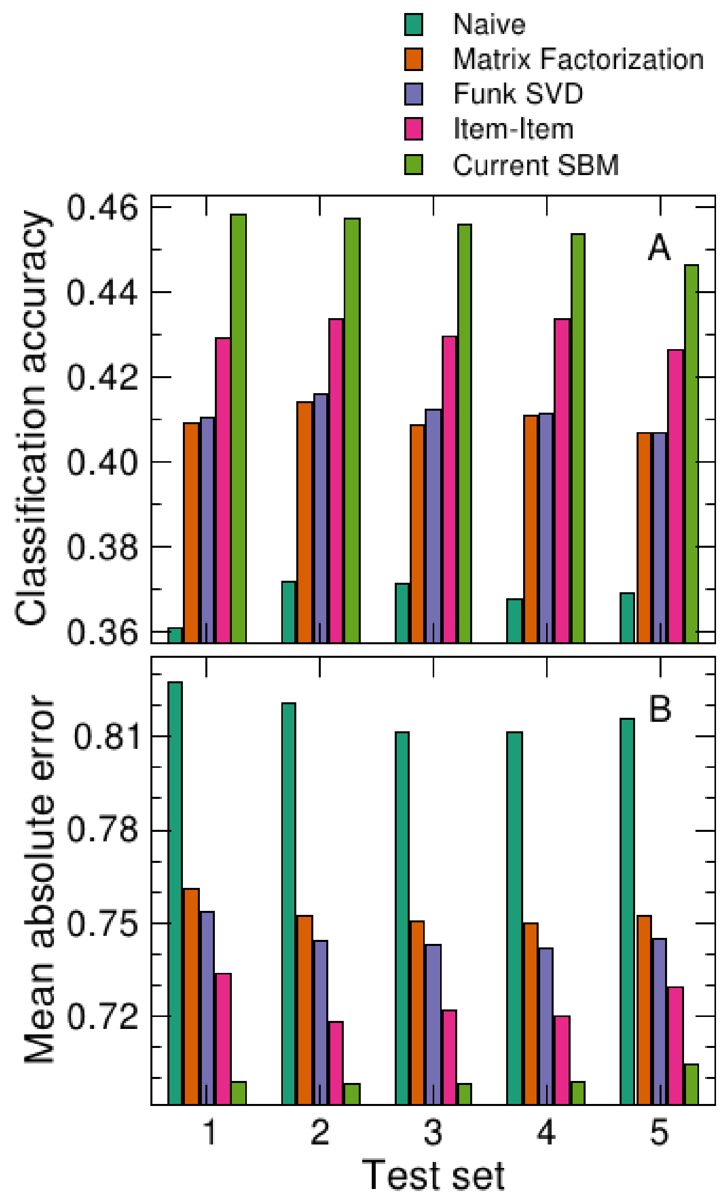
Our approach predicts human preferences better than state-of-the-art collaborative filtering algorithms

- ➔ MovieLens set: 100,000 real 1-5 movie ratings by ~1,000 users
- ➔ 5 independent splits of the data into 80,000 observed ratings and 20,000 validation ratings

Guimera, Llorente, Moro, Sales-Pardo (*PLOS ONE* 2012)

Our approach predicts human preferences better than state-of-the-art collaborative filtering algorithms

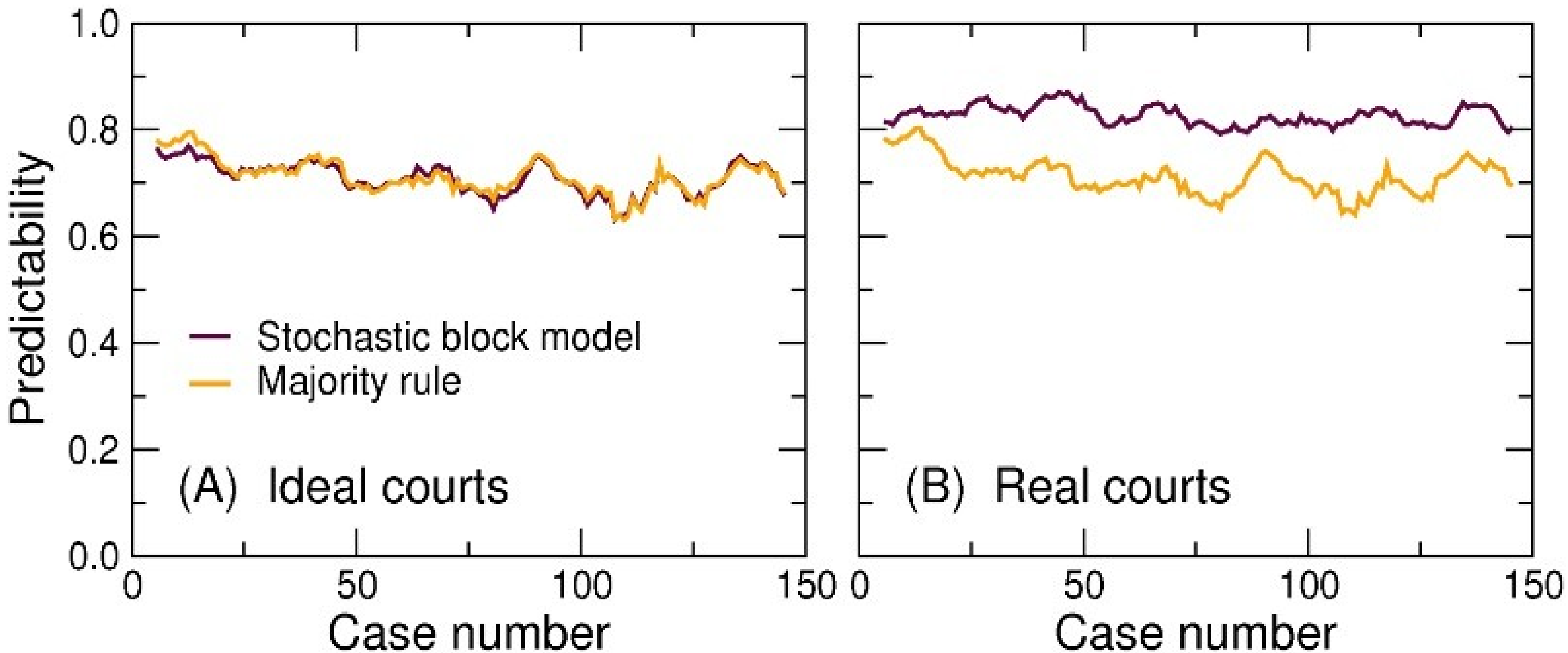
- MovieLens set: 100,000 real 1-5 movie ratings by ~1,000 users
- 5 independent splits of the data into 80,000 observed ratings and 20,000 validation ratings



Guimera, Llorente, Moro, Sales-Pardo (*PLOS ONE* 2012)

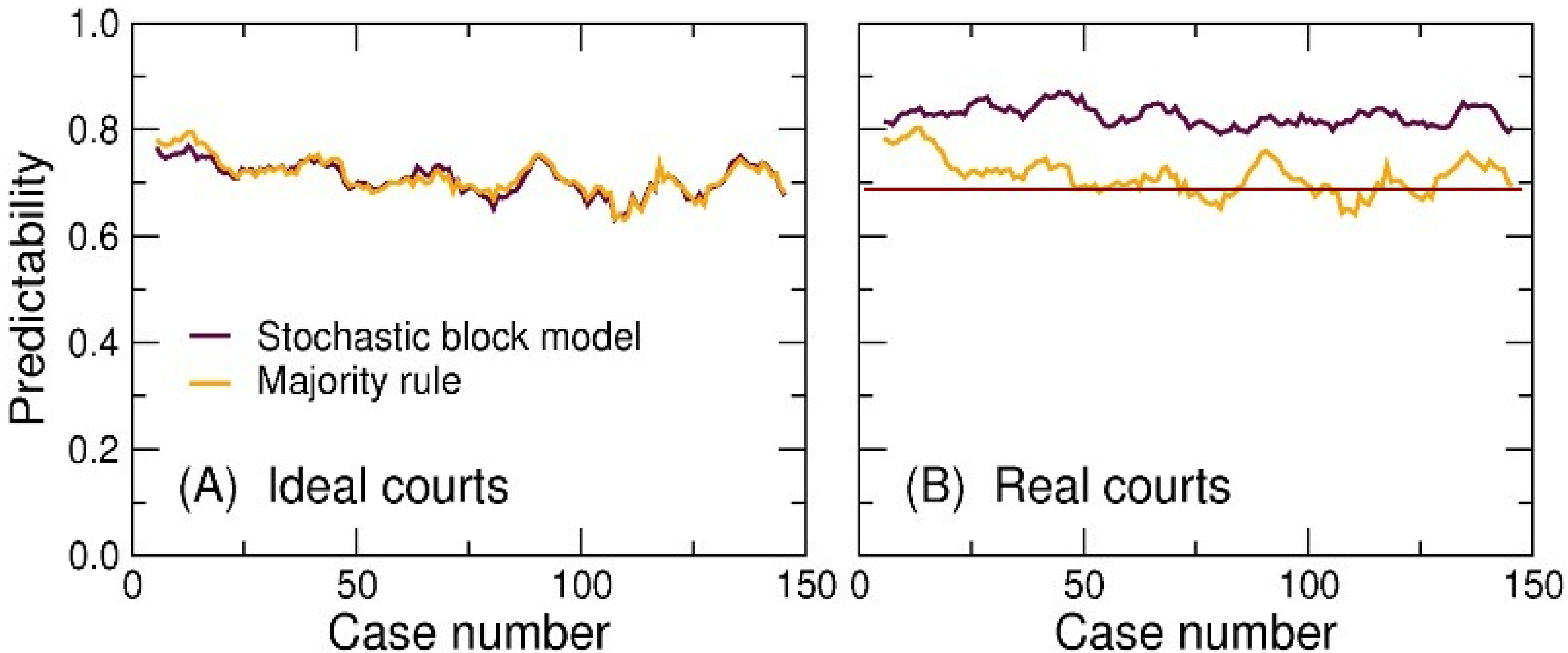
Can we predict what a US Supreme Court justice votes based on what the others did?

Supreme Court votes are more predictable than expected from ideal courts



Guimera, Sales-Pardo, *PLOS ONE* (2011)

Supreme Court votes are more predictable than expected from ideal courts

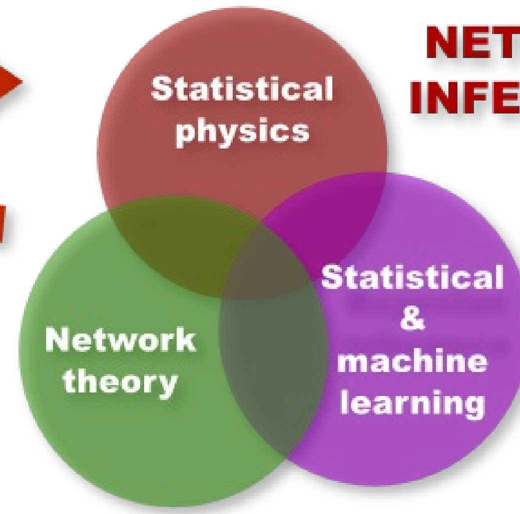


Guimera, Sales-Pardo, *PLOS ONE* (2011)

Outline

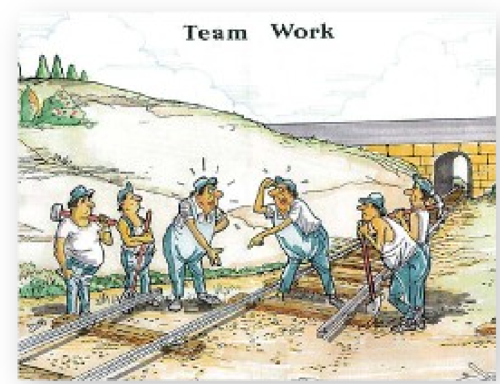
Can we help to clean up noisy network data?

NETWORK INFERENCE



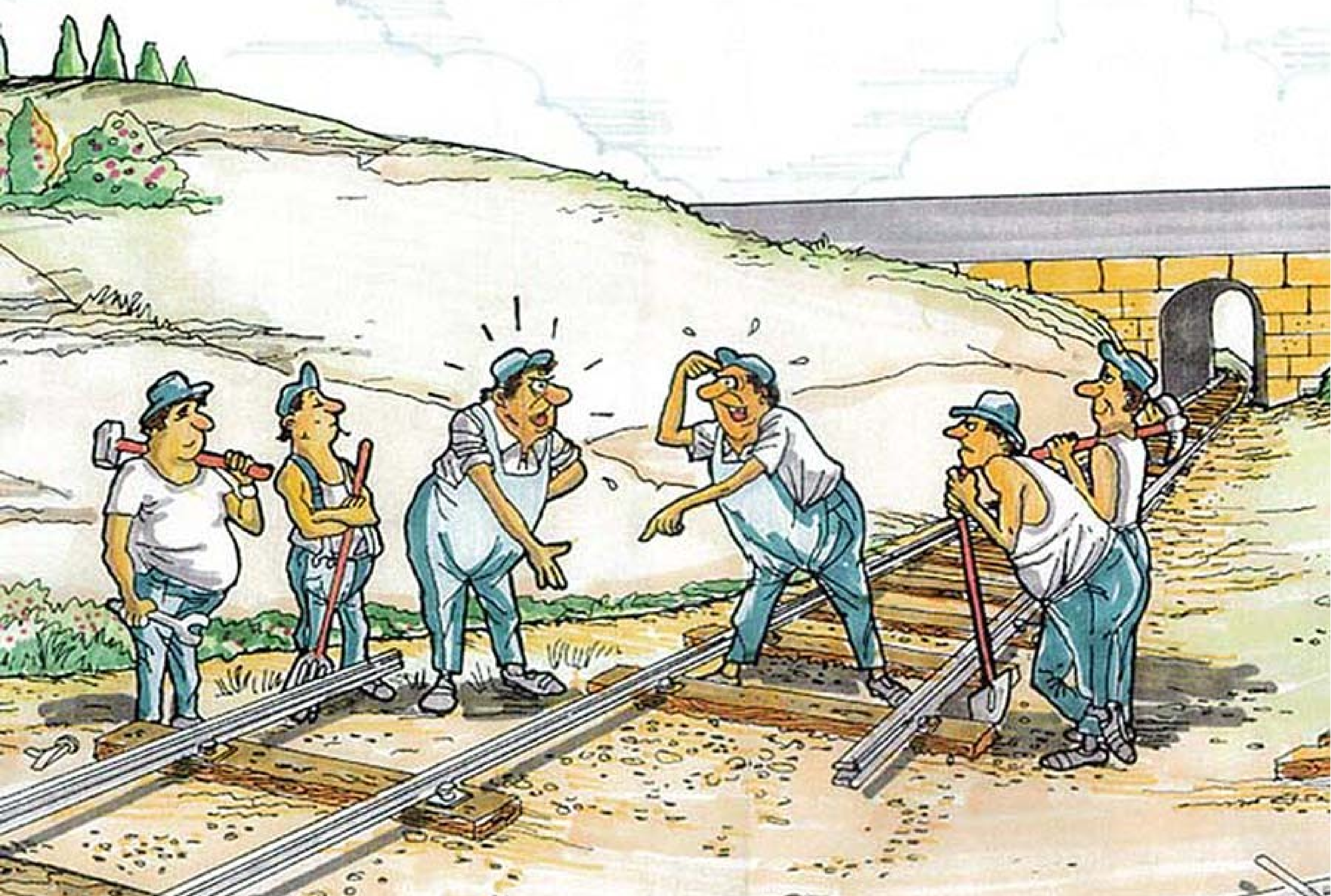
Can we uncover unknown drug interactions?

Can we predict human decisions?



Can we predict conflict in small teams?

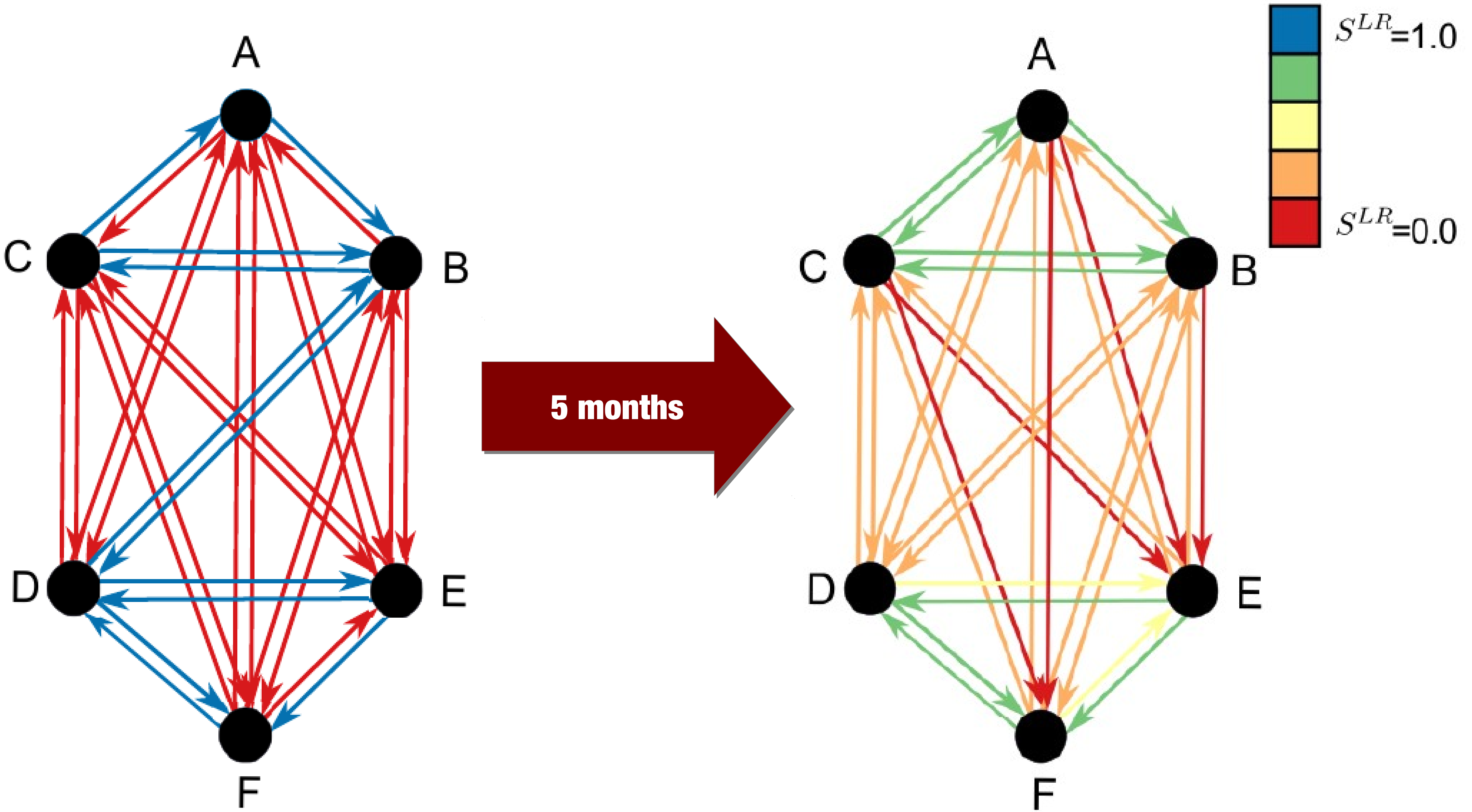
Team Work



Tracking team conflict in the real world

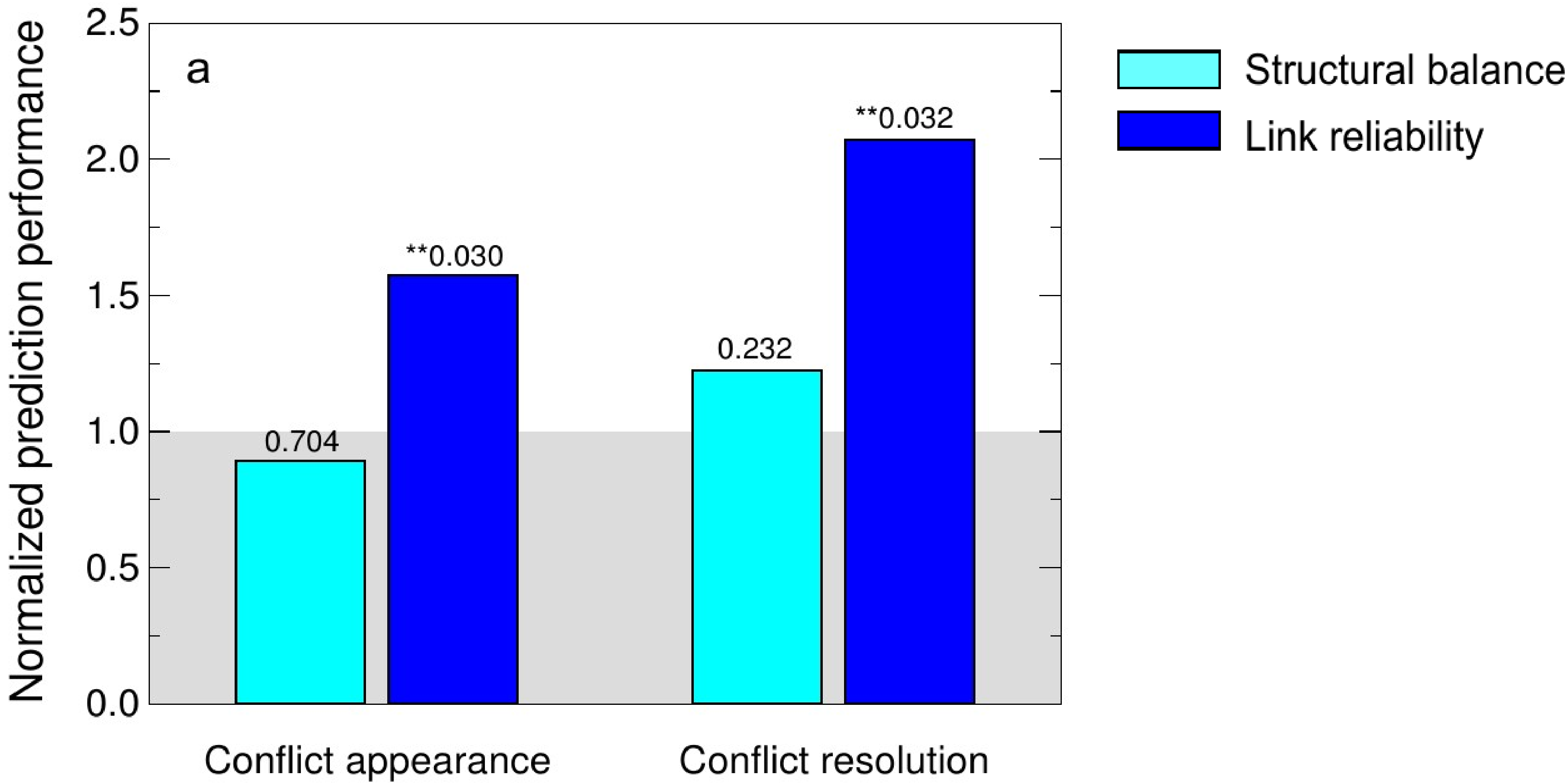
- 16 teams with ~6 people, working on a real project during 9 months
- We administer 2 surveys:
  - First: After 4 months working together
  - Second: At the end of the project
- “Would you like to work with this person again in the future”

Can we predict where conflict is going to arise and where it is going to resolve?



Rovira-Asenjo, Gumi, Sales-Pardo, Guimera, *in press* (2013)

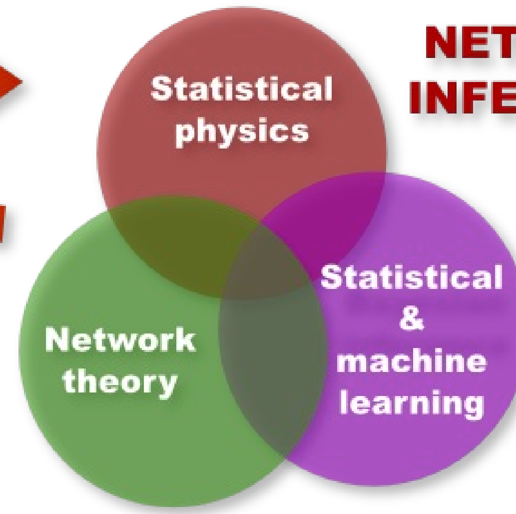
Our approach predicts conflict appearance and conflict resolution whereas structural balance does not



Rovira-Asenjo, Gumi, Sales-Pardo, Guimera, *in press* (2013)

Outline

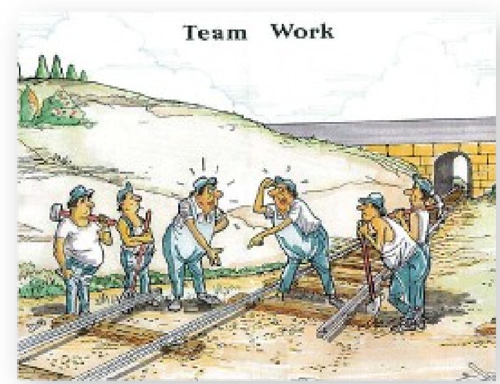
Can we help to clean up noisy network data?



NETWORK INFERENCE

Can we uncover unknown drug interactions?

Can we predict human decisions?



Can we predict conflict in small teams?

Thank you

- T. Gumí, A. Llorente, E. Moro, N. Rovira-Asenjo, M. Sales-Pardo
- Funding

*iCrea INSTITUCIÓ CATALANA DE RECERCA I ESTUDIS AVANÇATS

UNIVERSITAT ROVIRA I VIRGILI

MARIE CURIE ACTIONS

SEVENTH FRAMEWORK PROGRAMME

JAMES S. MCDONNELL FOUNDATION

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN

- More information:
  - http://seeslab.info
  - @sees_lab